Creating a Data Management Plan

March 2018

Stats4SD

Introduction

The main purpose of a data management plan is to help you produce high quality data thus reducing the risk of producing results that are not robust. A general aim of research is to improve the quality of individuals' lives; for example, by reducing poverty, promoting development, relieving pain, etc. Unreliable results are not useful for this purpose. By considering and detailing the data management tasks needed throughout the research project, you can ensure you have the necessary resources in place in terms of time, people skills, equipment, and finances. A data management checklist will help ensure that nothing is overlooked.

This document is aimed initially at the Principal Investigator (PI) whose task it is to allocate data management responsibilities to a member of the team. The person with data management responsibilities will be referred to as the "Data Manager".

Project Plan & Activity Protocol

In this document we will consider two levels – the project level and the activity level. At the project level the "Data Management Policy" aims to establish principles and agreements concerning data generated by the project as well as considering the resources that will be needed. At the activity level a "Data Management Plan" details the specific procedures that will be carried out to put the policy into operation. There should be a data management plan for each research activity. The PI has responsibility for drawing up and ensuring implementation of the overall policy while the Data Manager is responsible for activity level plans and procedures.

For example, it is CCAFS policy to archive data generated from project activities. This could be expressed in the CCAFS Data Management Policy as:

> *Data generated by CCAFS will be submitted to a public archive within 24 months of collection.*

Each activity level plan should then detail the steps that will be taken to prepare the data for archiving including anonymisation, producing the data dictionary, etc. The plan will also say who is responsible for each task.

Project Level Plan

The Project Data Management Plan would generally be drawn up by the PI. By the time of writing the plan, the PI should have a good idea of what activities are to be carried out and should therefore be able to make decisions about resources and the allocation of data management responsibilities.

Do I need a Data Manager?

One of the first decisions to be made is whether a data manager is needed for the project. The choice is basically between:

1. Having a specialist data manager to whom all data management responsibilities are allocated; or

2. Allocating data management responsibilities to scientists.

The decision depends on the size and complexity of the study and the skills of the scientists.

> *Most projects that involve a team of scientists (as opposed to a single researcher) are likely to need a data manager.*

We would recommend the inclusion of a data manager in most projects. This may correspond to the allocation of data management responsibilities to an existing member of the team who has the relevant skills, time and inclination to do the job well, or may involve recruiting a new member of staff with the relevant skills. A separate guide exists listing the terms of reference for a project data manager.

Principles & Agreements

As already mentioned the data management policy aims to establish principles and agreements with respect to data generated by the project. These should include:

- Data ownership
- Data sharing & access
- Ethics, privacy and copyright
- Archiving
- Quality Standards & Security
- Resources & Responsibilities

Data Ownership

Data ownership is often a contentious issue and it is important to draw up agreements from the outset to avoid problems later. All researchers will need to sign up to these agreements. Data ownership is covered in more detail in a separate guide. As previously mentioned, CCAFS is now governed by the CGIAR Open Access Policy and the principles of this policy should be made clear to all researchers working on CCAFS projects so that they understand from the outset that they cannot keep the data to themselves.

Data Sharing & Access

This is linked to data ownership in that if it is established that all data are owned by the project, it follows that all team members must have access to the data. However, there are issues regarding confidentiality, so it might be that only a limited few have access to the raw data, but others can access the anonymised data, or that data access occurs in a staged manner with more individuals being given access with time. Thus, the decisions to be made are:

- Who has access?
- Who grants access?
- How is data accessed?

Ethics, Privacy & Copyright

If the research involves collecting data on individuals the researcher would generally need to obtain ethical approval or establish the code by which the project will work; respondents must be fully informed about the purpose of the study; their data cannot be included unless they have given

informed consent and they must be given the option of withdrawing from the project at any time; all personal data must remain confidential – in general individuals should not be identifiable from any data that is put into the public domain although there may be exceptions; GIS references pose a challenge to this. Copies of the consent forms should be included in the project archive.

If any copyrighted data or data collection tools are to be used, then permission must be sought from the copyright holder. Any legal restrictions should be considered.

Archiving

This is likely to be a broad statement in the data management policy stating for example that all data generated by the project will be put into the public domain within a specified timeframe. It would then be the responsibility of the data manager to draw up a document describing the requirements for archiving which scientists must incorporate into the data management plans for their activities.

Quality Assurance & Security

Where quality assurance and security are concerned the principles stated in the policy document should prompt the data manager to include details under these headers in the activity level plans. For example, the project policy document might include the statement:

> *Project data will be subject to a quality assurance process*

The activity data management plans should then detail the steps to be taken. For example:

- CSPro will be used for data entry and a customised data entry system will be produced;
- Automatic skips will be programmed into the system which follow the skips in the questionnaire;
- Double-data entry will be used and <the data manager> will carry out the data entry comparisons and subsequent corrections;
- Backups to external hard drives will be taken regularly with incremental backups taken at the end of each day and full backups once a week. <The data manager> is responsible for ensuring that these backups are done;
- Etc.

Resources & Responsibilities

Resources include people (skills & responsibilities), equipment (hardware & software), time and money. The PI should be aware of the resources currently available and, by examining the list of planned research activities, should be able to draw up a list of requirements adding in some element for contingency. The plan should clearly state who is responsible for the data management of the project.

In the box below, we have set out a draft template for a project data management plan.

Data Management Policy for <Project Name>

This plan sets out the data management principles and responsibilities for the <Project Name> research project. For each research activity a data management plan must be produced. These plans should detail the steps that will be followed to ensure these principles are adhered to.

Data Manager: The data manager for the project will be <name of DM>. He/she will have overall responsibility for ensuring the timely completion of all data management tasks.

Data Ownership: All data generated by the project will be owned jointly by the collaborating institutions <names of collaborators>. A data ownership agreement will be drawn up and signed by all parties.

Data Sharing & Access: All scientists in the project team must have access to the data as and when needed.

Ethics, Privacy & Copyright: Ethical approval must be sought for research activities that involve human subjects. Confidentiality of personal data must be maintained. Any copyrighted materials used in the research must be acknowledged correctly.

Quality Standards & Security: All project data will be subject to a quality assurance process. Regular backups will be taken throughout the project.

Archiving: All data generated by the project will be archived within 24 months of data collection

Activity Level Protocol

The activity level data management plan would be drawn up by the data manager or the person with data management responsibilities within the team. This is a much more detailed document explaining how the principles are going to be achieved. For example, how they intend to set up a data and document storage facility for data sharing among the team.

The plan will naturally vary according to the type of activity but would generally include the following elements:

Data Collection

Capture Methods

Briefly describe the activity.

- Will you be running survey(s) or conducting experiment(s)/trial(s) to collect your data? Describe the methods to be used.
- What technology will be used for capturing the data – paper, mobile device, etc.?

Data Description

Include a brief description of the information to be gathered including the nature, scope and scale of the data that will be generated. For example:

- What are the study units and how many will there be? For example: we will be collecting data on 20 households in each of seven villages.

- What mechanisms do you have to check that the correct amount of data is collected – i.e. that you have the right number of study units?
- Where appropriate have you determined the units of measurement to be used? How will you ensure consistency in the units used?

Secondary Data

Is there a need to review secondary data in your project activity? If yes, then describe the data to be used.

- Where are these data currently stored?
- How are they to be accessed?
- Who owns the IPR on these data?
- Can new data generated by the current activity be easily linked to the secondary data – what are the linking fields?

Computerisation & Storage

Data Entry

- How will you capture the data – on paper or directly onto hand-held devices or laptops?
- What software will you be using for data entry?
- Have you a customised data entry system or will you be preparing one? What data checks are/will be included in the system?
- Will the data entry system be documented?
- If recording directly onto hand-held devices what mechanisms do you have to ensure quality? For example: two researchers will be present during data collection to validate values.
- Have data entry staff been trained in the use of the data entry system –how long was spent on the training – do you have any mechanisms for checking competency of data entry staff?
- Are you using double data entry (DDE) – if yes, then who will carry out the data comparisons and how will this be done?

Quality Assurance

How do you intend to ensure that your data are of high quality? Detail the data checks you are intending to carry out with a comprehensive checklist of consistency checks – for example, if a farmer only has access to one acre of land then he/she cannot be using more than one acre for growing crops; harvest date cannot be before planting date; etc.

Keep an audit trail detailing problems and inconsistencies found in the data and what was done about them. Include this information as part of the data quality assessment document.

Data Structure & Organisation

Describe the structure of the data; in particular how many levels you are expecting and what they are; for example, village level, household level, and individual level. How many records/cases are expected at each level or is this variable. This should match with the expected number of study units from your sampling scheme.

- What field(s) will be used to link the data at the different levels?

- What formats are to be used for storing the data? Note the format for data entry may differ to that used for storage and archiving. For example, data entry might be done in CSPro but exported to SPSS for storage, analysis and archiving.

Data Dictionary

The data dictionary should include the following information for each variable:

- **Name**– variable names should be kept short, ideally no more than 8 to 10 characters, this makes them easier to use when programming. Do you have a naming convention for your variables?
- **Label –** the label gives an indication of what the data represents. For example: name=RESPAGE, label = "Age of respondent in whole years"
- **Codes and labels –** If numeric or short text codes are used then the dictionary should detail these for each variable. For example, 1=Male, 2=Female.
- **Missing value codes** – it is useful to include codes for missing values. This should always be a value that is not feasible for the variable. For instance, a missing value code of "99" would be suitable for coded data where there are only 20 valid codes (01 to 20) but would not be suitable for "Age of respondent" as it is possible that the respondent is 99yrs of age. You might also want to distinguish between different types of missing value. For example, "not applicable", "refused to answer", etc.
- **Unit of measurement** – where relevant make sure you include the unit of measurement used for the data.

The data dictionary should also indicate the unique identifiers or primary key fields. Specify the number of records and the number of variables.

- Will you be deriving any variables such as a standard set of indices? If yes, describe these derived variables and explain how they are calculated. Include the syntax for creating these variables.

Storage & Backup

Describe how you intend to make the data available to all members of the project team and how you will keep others informed about updates.

- Will you be using a DDS (Data and Document Storage Facility)? If yes, describe the system to be used.
- Will you have a file and folder naming convention? If yes, say who is responsible for ensuring this is followed?
- How will you manage the DDS to avoid accidental deletions? How will you keep it organised? Will it be password-protected? Will all team members have Read/Write access or only the data manager?
- Will you keep different versions of files? If yes, how will you distinguish between them – e.g. by including the date in the filename?

Detail your backup procedures.

- What mechanisms do you have for ensuring the security of your data?

- How often do you take backups?
- Who is responsible for backing up the data and documents?
- Which files are included in your backup?
- Where are the backups stored?
- Have you tested your restore method?

Don't rely on others to do your backups!

Legal Aspects

Ethics & Privacy

- Are you collecting any personal data in this research activity?
- Have you obtained ethical approval – what organisation granted this approval?
- Have all fieldworkers been adequately trained to be able to explain the consent process to potential respondents?
- Have you prepared an information sheet for respondents together with a consent form?
- What mechanisms do you have for ensuring the confidentiality of personal data?
- Detail steps taken to anonymise the data

Data Ownership

Have all team members been made aware and agreed to the terms of the Project Data Ownership agreement?

Copyrighted Material

Are you using any copyrighted data collection tools or methods in your research activity? If yes, have you sought permission from the copyright holder? Include here details of the copyrighted material together with any legal restrictions which might impact on how the data are used.

Archiving & Preservation

The decision on whether to archive data is taken at the project level. The decision on where to archive might be at the project level or at the activity level.

- Where will the data be archived?
- Will there be any restrictions on access to the archive?

For further information on archiving please see the separate guide "Principles for Archiving and Sharing Data".

Training & Responsibilities

Name the individuals responsible for ensuring these tasks are carried out. This would normally be the data manager but might also involve scientists within the team.

Describe how you are intending to make team members aware of their responsibilities and the requirements of data management. For example, are you going to have seminars or training events or produce documented guidance?

Do you have a time scheduled for training the data entry staff? How long will this training take?

Summary

So, to summarise there should be a policy document at the project level created by the PI which details the key data management principles. Then, for each activity, there should be a data management plan which details the steps you intend to follow to put these principles into action. This would normally be done by the project data manager. Depending on the size and scope of the activity he/she may well liaise with the team of scientists for the activity perhaps delegating from the data management tasks. Thus, we have:

- Project Level Policy – general principles of what you intend to do
- Activity Level Plan – details of how you intend to do it

External Resources

- IHSN – Principles and Good Practice for Preserving Data
- ICPSR – Guidelines for Effective Data Management Plans

Associate Videos

Videos accompanying the original release of the CCAFS Data Management Support Pack in 2013 are available as a playlist on the Statistical Services Centre YouTube Channel at https://www.youtube.com/channel/UCs7EU95YMjhvNozJKCD92xQ/playlists. These videos have not been updated since the original release but are mostly still relevant.

In particular the playlist includes a video on Data Management Plans available from the following link:
https://www.youtube.com/watch?v=Q8jX_cH0C60&index=3&list=PLK5PktXR1tmNRaUPsFiYlyhg2Iui0xgpj